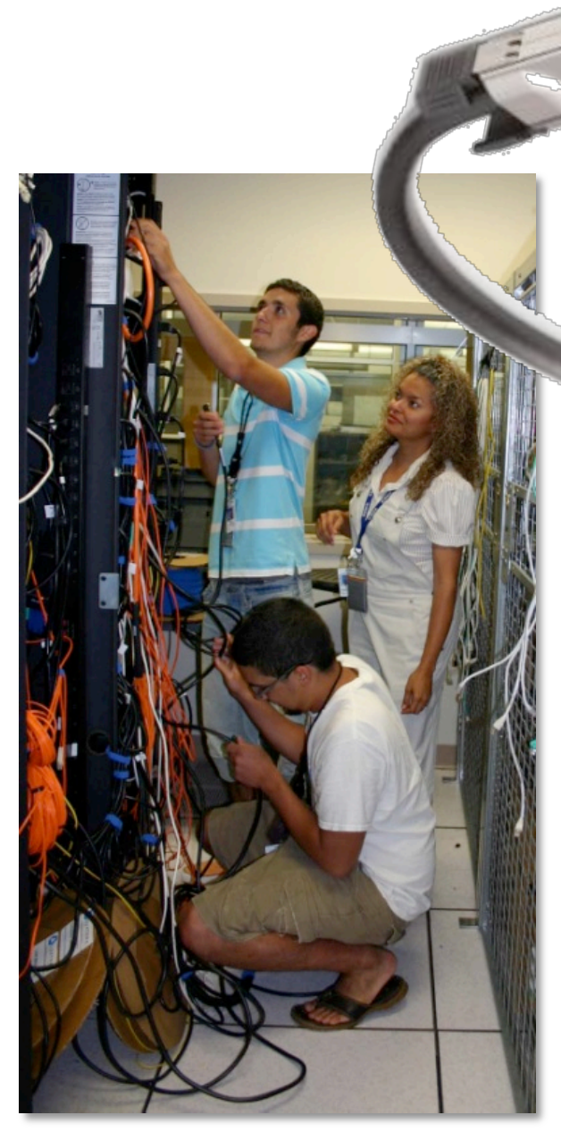
Performance Analysis and Evaluation of LANL's PaScalBB IO nodes using PCIe Gen 2.0 Quad-Data-Rate Infiniband and Multiple 10-Gigabit Ethernets

Abstract

I/O nodes are the key components used in LANL's PaScalBB (Parallel Scalable Back Bone) infrastructure to carry data traffic between backend compute nodes and global scratch file systems. Combining Infiniband Quad-Data Rate (QDR) HCA with multiple 10-Gigabit IPC Ethernet links can potentially alleviate currently existing I/O traffic bottlenecks. In this experiment we set up a small-scale PaScalBB test bed and conduct a sequence of I/O node performance tests. The purpose of these experiments is to find an enhanced network configuration that can be applied to LANL's future supercomputer utilizing PaScalBB architecture.

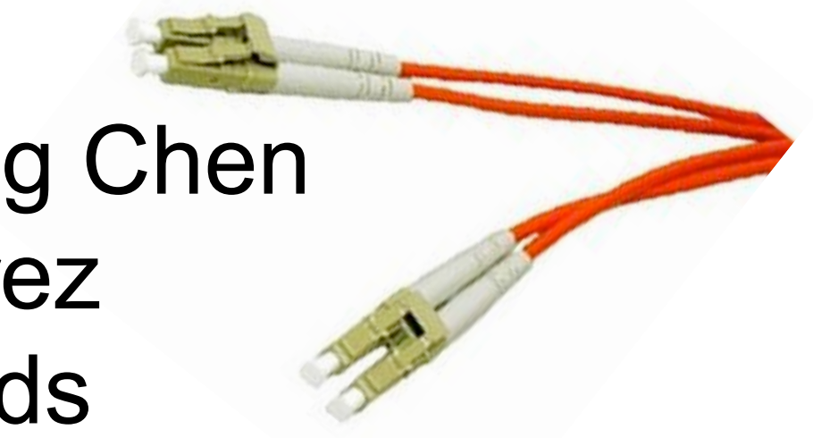


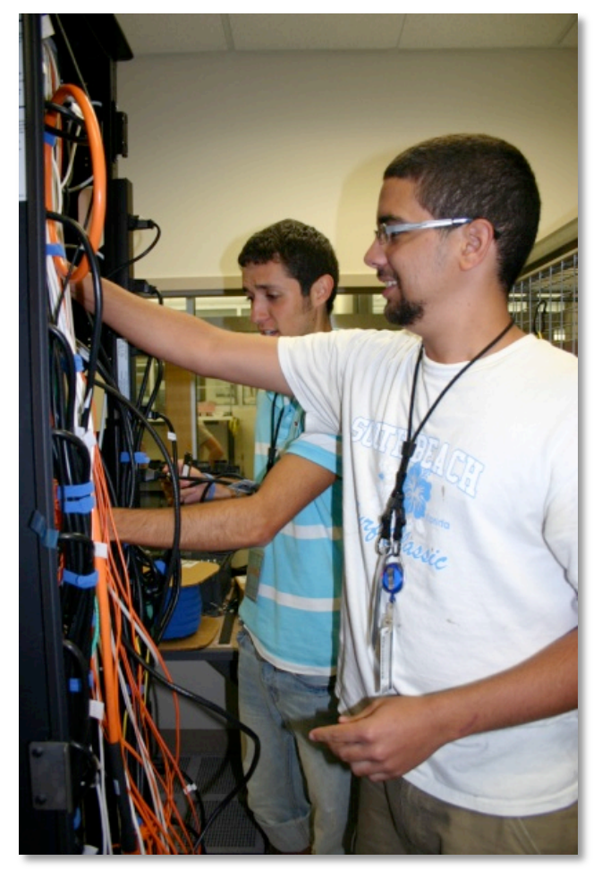
Checking the cluster connections

Authors: Juan C. Franco (Univ. of N. Texas)
Rocio Perez-Medina (NNMC)
Daniel L. Illescas (UNM)

Mentors: Hsing-bung Chen
Alfred Torrez
Parks Fields

Instructor: Andree Jacobson



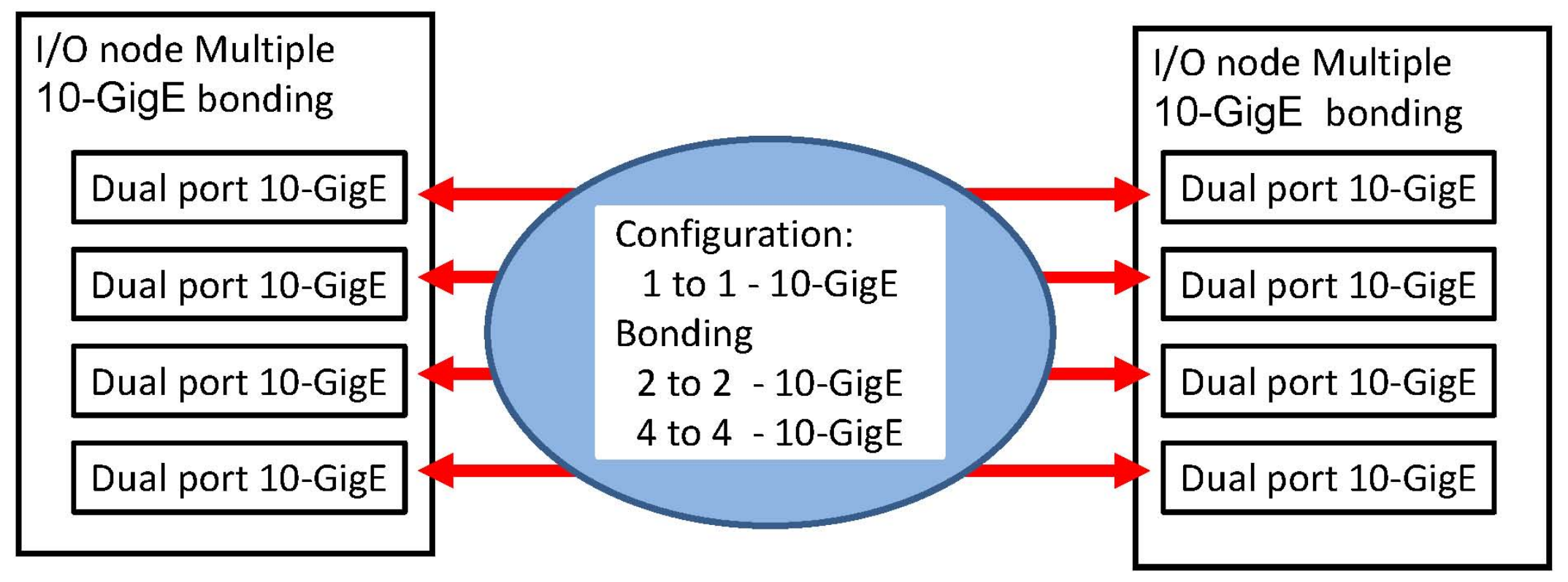


Doing the IB/QDR and 10-Gigabit bonding

ConnectX EN
Single/DualPort 10-Gigabit Ethernet Adapters with PCI Express

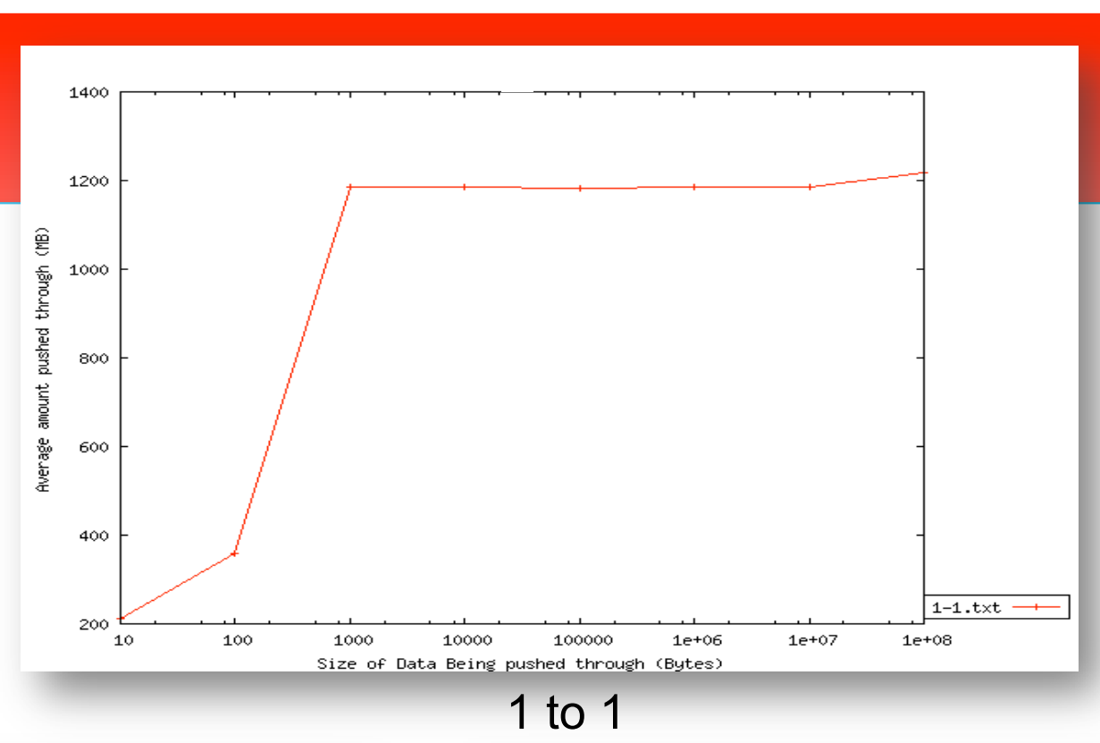
Testing the bonding

Performance Back-to-Back Multiple 10-Gigabit bonding testing

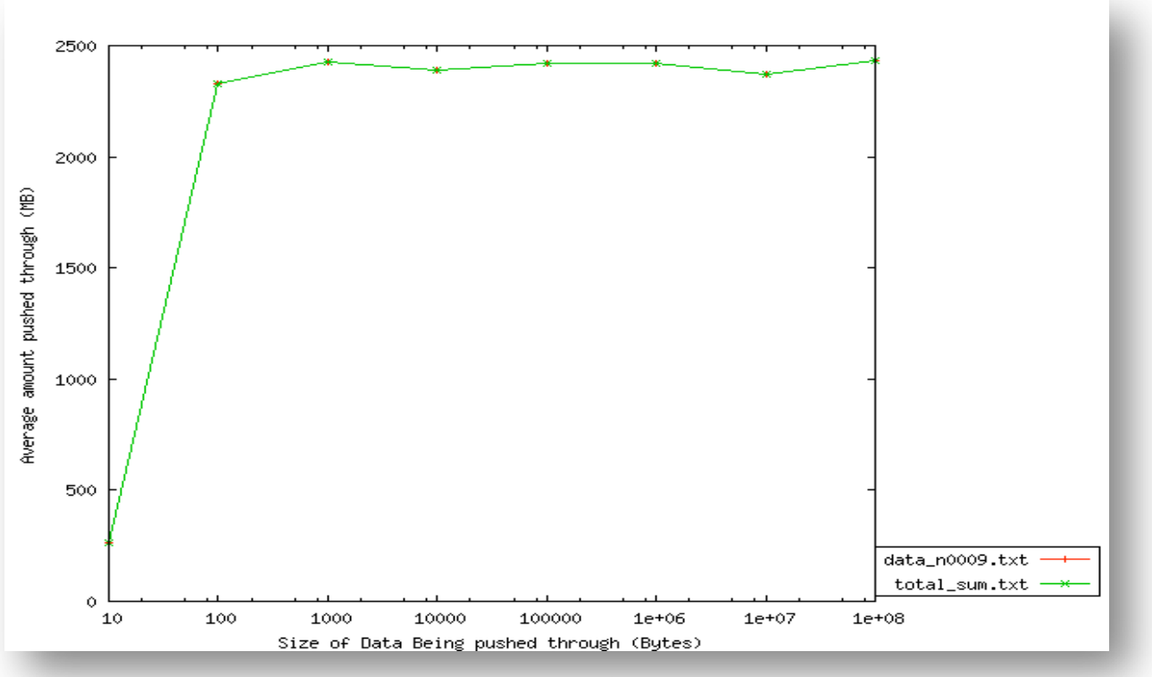


Arista 10-Gigabit transceivers.
Set bonding Mode 0 and Mode 5 to get load balancing.
Increased TCP buffer size, read and write memory to 16.7 MB.
Set the TCP timestamp and TCP sack to 1
The changes we made to the TCP stack increased the bandwidth performance by 10x from the TCP default setting. By default the setting are set too small.

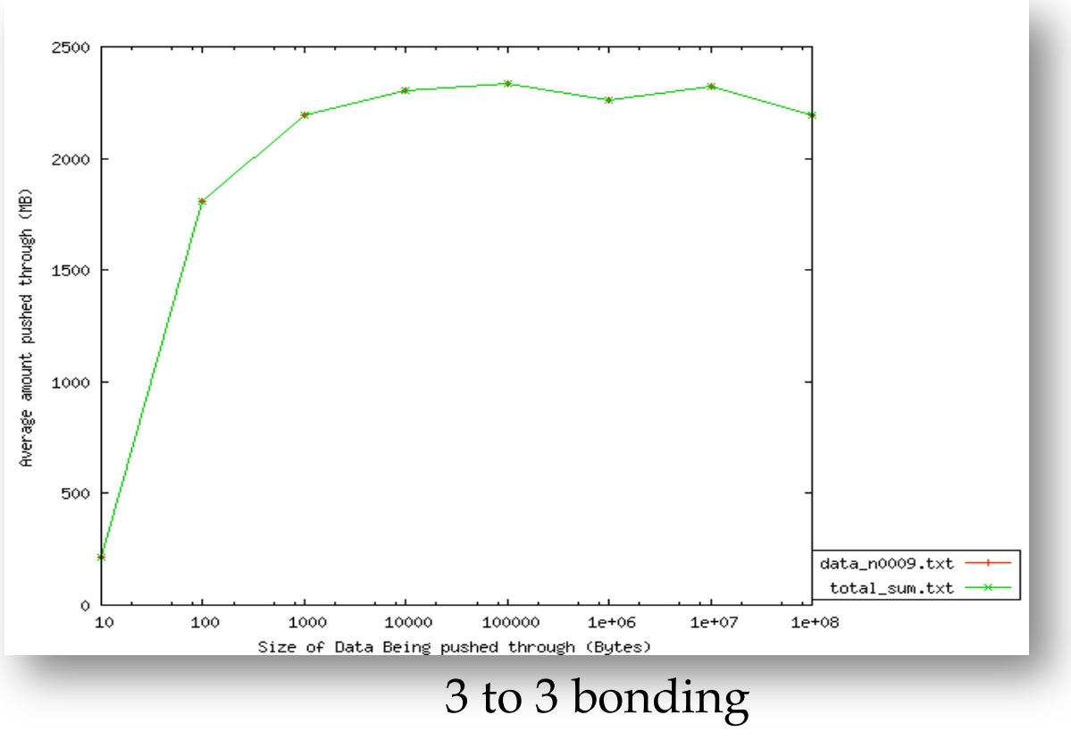
Actual results:
1 to 1 no bonding - 9.5 gigabit per second
3 to 3 bonding - 18.3 gigabit per second
4 to 4 bonding - 19.2 gigabit per second



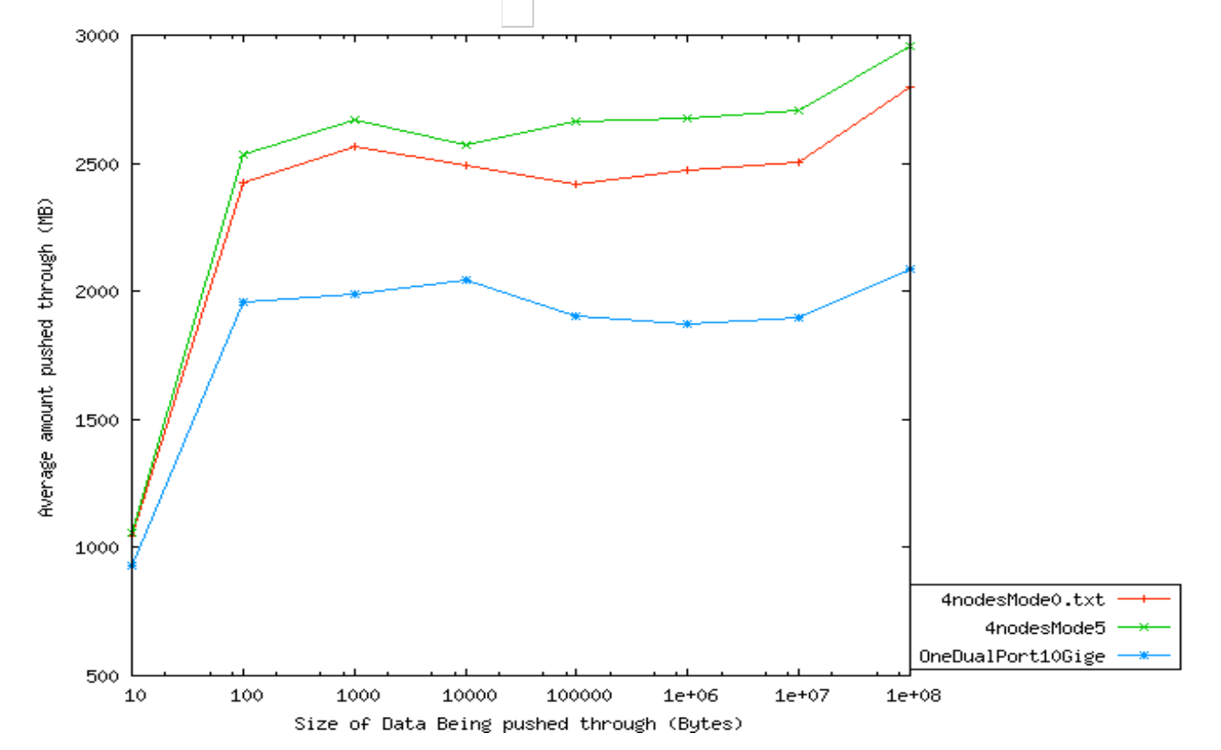
1 to 1



4 to 4 bonding



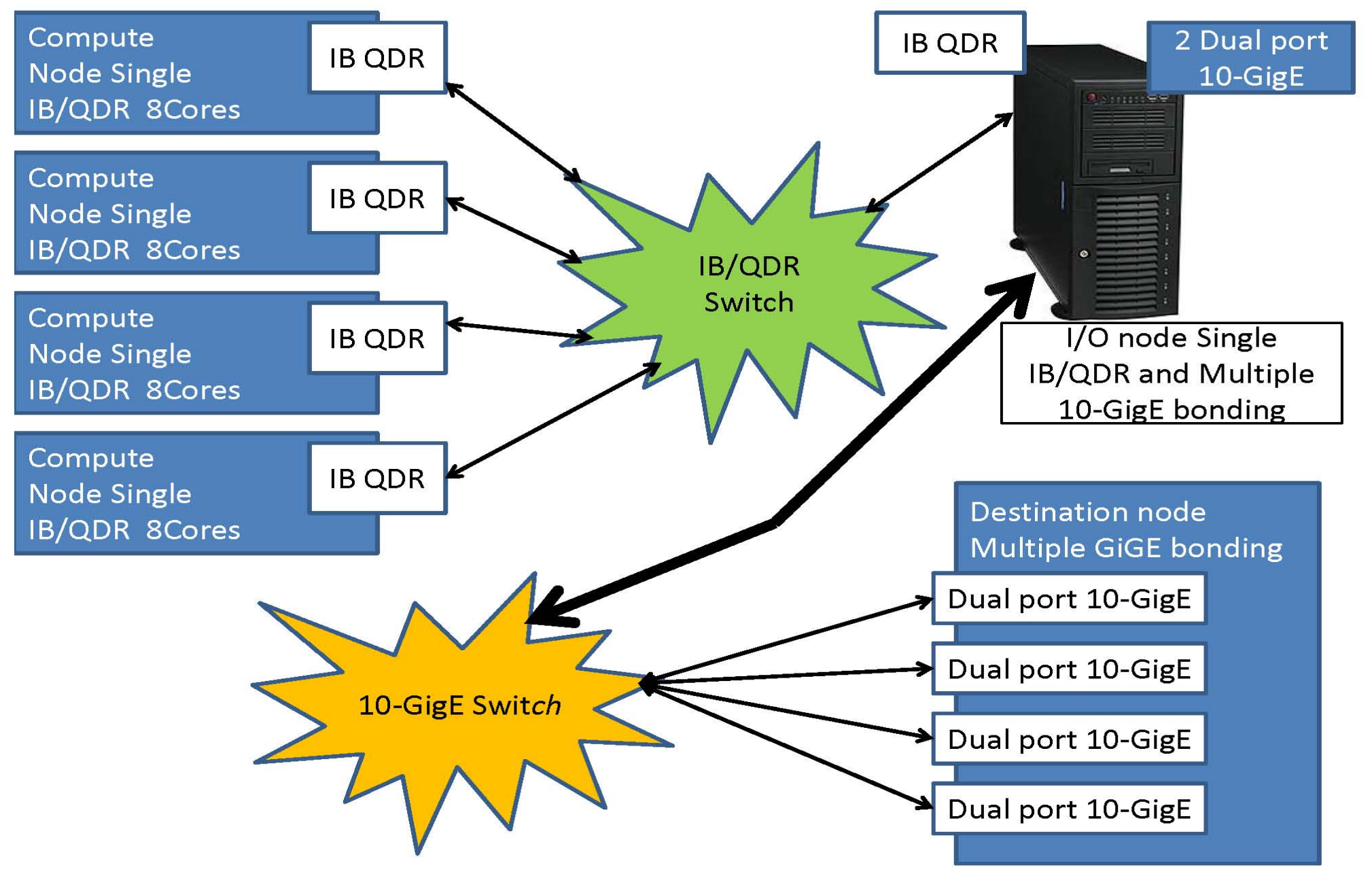
3 to 3 bonding



Mode 0 (Round Robin) and Mode 5 (Load Sharing) Performance

ConnectX
Single/Dual-Port Infiniband Adapter cards with PCI Express 2.0

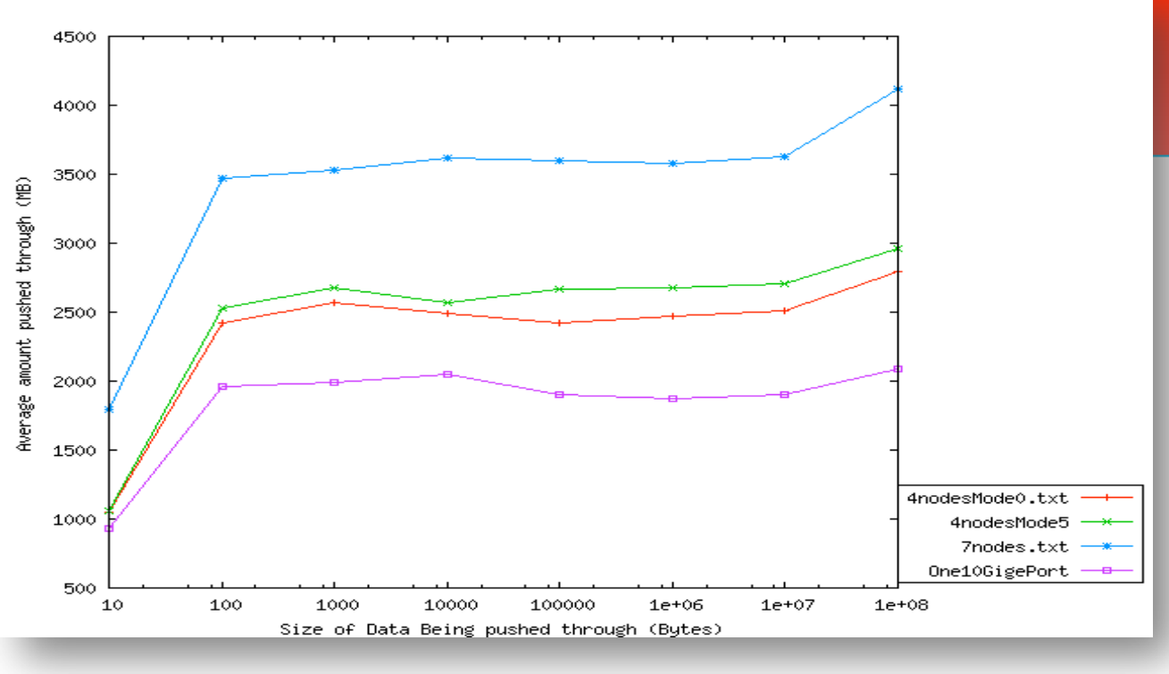
IB/QDR + Multiple 10-Gigabit bonding


Compute Node Single IB/QDR 8Cores
IB QDR
Compute Node Single IB/QDR 8Cores
IB QDR
Compute Node Single IB/QDR 8Cores
IB QDR
Compute Node Single IB/QDR 8Cores
IB QDR
IB/QDR Switch
IB QDR
2 Dual port 10-GigE
I/O node Single IB/QDR and Multiple 10-GigE bonding
10-GigE Switch
Destination node Multiple GiGE bonding
Dual port 10-GigE
Dual port 10-GigE
Dual port 10-GigE
Dual port 10-GigE

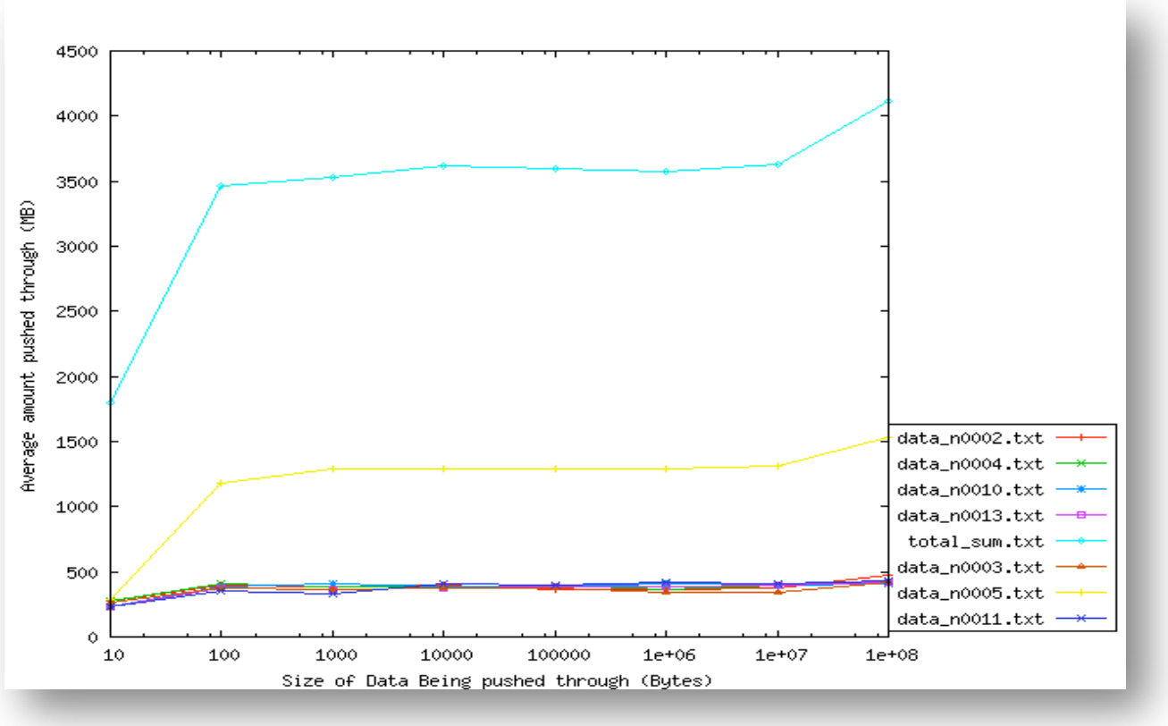
Set MTU to 65520 on the infiniband configuration.
Bond four 10-gigabit ports together.

Theoretical bandwidth performance – 40 gigabits per second
Physical bandwidth performance – 32 gigabits per second

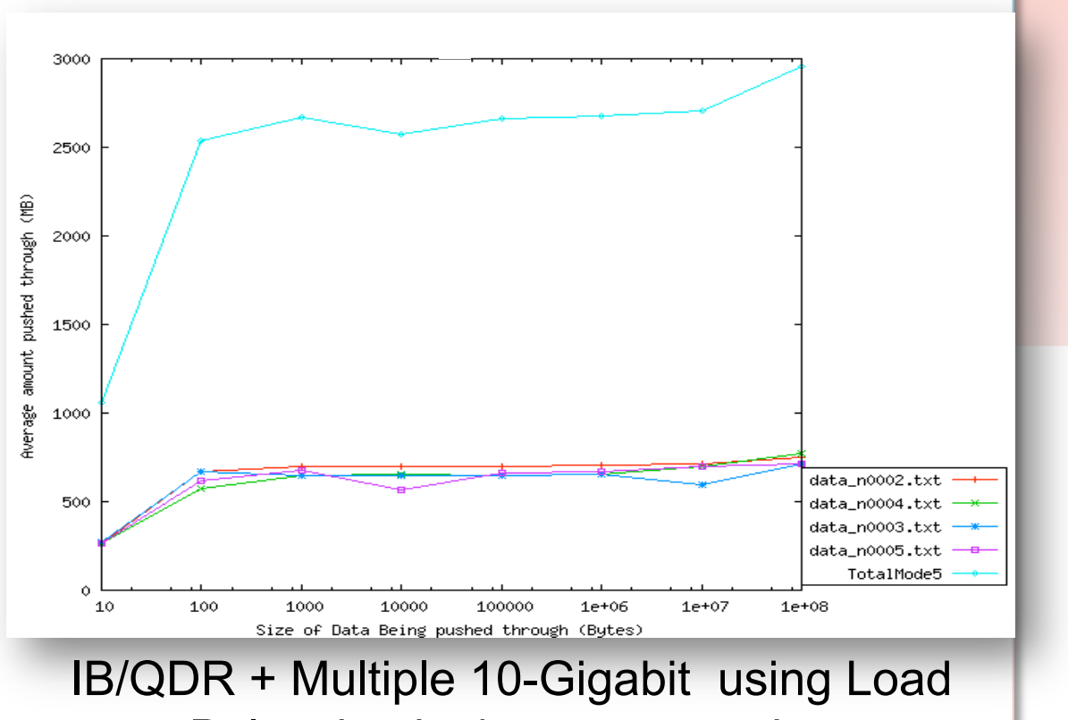
Actual results:
4 compute nodes – 24 gigabits per second
8 compute nodes – 28 gigabits per second



7 compute nodes , 4 compute nodes, and Dual Port 10-Gigabit card using Load Balancing



IB/QDR + Multiple 10-Gigabit using Load Balancing in 7 compute nodes



IB/QDR + Multiple 10-Gigabit using Load Balancing in 4 compute nodes

Conclusion

Our testing showed the bonding of 10-Gigabit ports will increase the bandwidth through a single I/O node and decrease the number of I/O nodes needed. This new configuration will enhance the performance and lower the cost of the PaScalBB infrastructure.

Los Alamos NATIONAL LABORATORY EST. 1943

National Security Education Center
A CONSORTIUM OF LANL INSTITUTES

ISTI
INFORMATION SCIENCE & TECHNOLOGY INSTITUTE

New Mexico Consortium's
IAS
Institute for Advanced Studies
at Los Alamos National Laboratory